**Multiple Linear Regressions – Module -7**

1. Prepare a prediction model for profit of 50\_startups data. Do transformations for getting better predictions of profit and make a table containing R^2 value for each prepared model.

**Business** **Problem**: Predicting the model for better Profits of 50 Stratups.

**Data Pre-processing:**

Predicting value (Y) is Profit and it is Continuous Data.

Input data are 4 features.

R.D.Spend : Data is Continuous

Administration : Data is Continuous

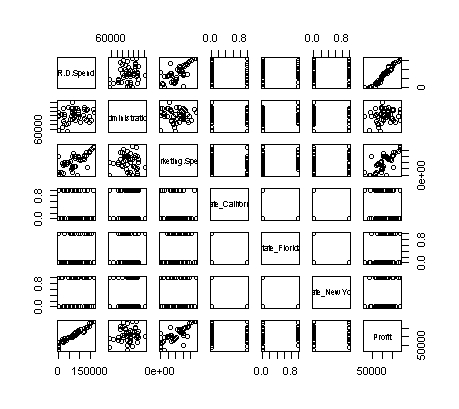
Marketing.Spend : Data is Continuous

State: Discrete in Categorical format, it California, Florida and New York. In order to process the state converting it into Dummy data.

Hence Y is continuous and X has multiple features, I’m going with multiple linear regression models.

**Model Building & Checking the Correlation:**

Plotting the Scatter Diagram:

By looking the diagram we came to know Administration field is not correlated with the Profit. Scatter diagram is subjective so taking the correlation value.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **R.D.Spend** | **Administration** | **Marketing.Spend** | **State\_California** | **State\_Florida** | **State\_New York** | **Profit** |
| **R.D.Spend** | 1 | 0.241955245 | 0.72424813 | -0.14316522 | 0.10571106 | 0.039068162 | 0.9729005 |
| **Administration** | 0.24195525 | 1 | -0.03215388 | -0.01547811 | 0.01049309 | 0.005145226 | 0.2007166 |
| **Marketing.Spend** | 0.72424813 | -0.032153875 | 1 | -0.16887523 | 0.20568545 | -0.0336698 | 0.7477657 |
| **State\_California** | -0.14316522 | -0.015478106 | -0.16887523 | 1 | -0.49236596 | -0.515151515 | -0.145837 |
| **State\_Florida** | 0.10571106 | 0.010493089 | 0.20568545 | -0.49236596 | 1 | -0.492365964 | 0.1162443 |
| **State\_New York** | 0.03906816 | 0.005145226 | -0.0336698 | -0.51515152 | -0.49236596 | 1 | 0.0313676 |
| **Profit** | 0.97290047 | 0.200716568 | 0.74776572 | -0.14583704 | 0.11624426 | 0.0313676 | 1 |

State fields are not correlated with Profit because it is categorical data. It has data in the following ways.

State Yes for 1

State No for 0

Mention the same in the Model.

Build the Model: model.strtup <- lm(Profit~Spend+Admin+Market+Cali+Florida)

###Not adding the NewYork – Due to degree of Freedom

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.012e+04 6.572e+03 7.626 1.06e-09 \*\*\*

Spend 8.057e-01 4.515e-02 17.846 < 2e-16 \*\*\*

Admin -2.682e-02 5.103e-02 -0.526 0.602

Market 2.723e-02 1.645e-02 1.655 0.105

Cali 4.189e+01 3.256e+03 0.013 0.990

Florida 2.407e+02 3.339e+03 0.072 0.943

Multiple R-squared: 0.9508, Adjusted R-squared: 0.9452

First Kept side the state related data because it is Categorical data. We have got good R-Squared value but there is P value is higher than accepted value (0.05) for Administration and Market spending.

Probability value is very high to state features, So they are not relevant to Profit. We will confirm by them using avPlot.

**Check the Co linearity Problem:**

Administration correlation co-efficient (r) has very less. So there will be chance for co linearity.

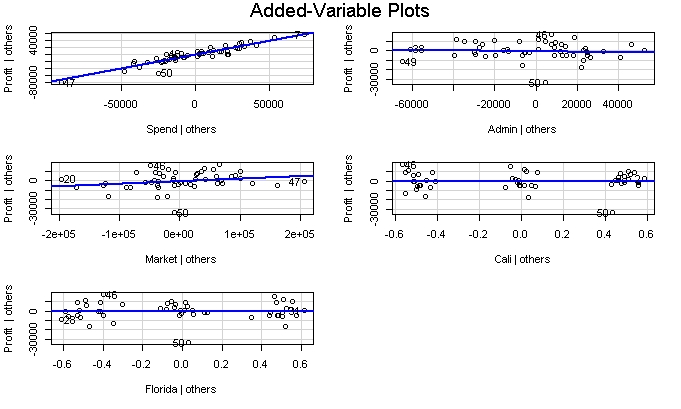
When we build the models individually for Administration and Market Spend, Administration is showing some dependencies ( P- value less ) on Market Spend.

Variance Inflation factor to check co linearity between the variables:

Spend Admin Market Cali Florida

2.495511 1.177766 2.416797 1.335061 1.361299

By looking the VIF data, not able to differentiate the exact collinearity variable. So draw the avPlot.



From the avPlot, we are sure Administration has co linearity with Market Spend and States are not relevant field.

Building the model by removing the Administration feature from dataset.

model1.strtup <- lm(Profit~Spend+Market)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.694e+04 3.343e+03 14.043 <2e-16 \*\*\*

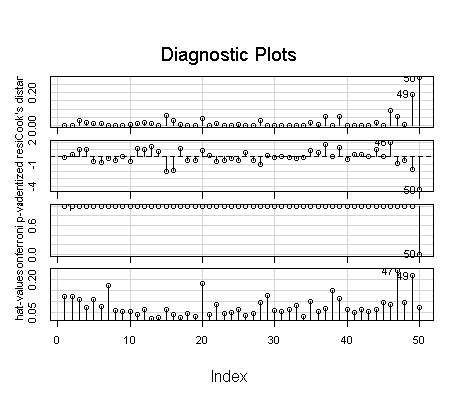
Spend 7.967e-01 4.245e-02 18.771 <2e-16 \*\*\*

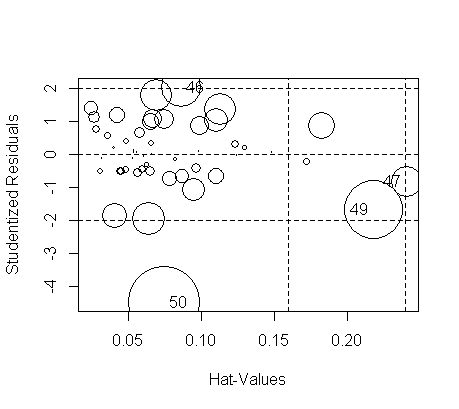
Market 2.975e-02 1.615e-02 1.842 0.072 .

Multiple R-squared: 0.9505, Adjusted R-squared: 0.946

Still R-squared value is good but P-value of Market is higher than the accepted. Something we need to look into it.

**Checking for influential observation:** When execute the Influential Index Plot & Influential Plot.





StudRes Hat CookD

46 2.0357210 0.1277290 0.09439478

47 -0.8354542 0.2654200 0.04232333

49 -1.6860294 0.2558868 0.15637613

50 -4.4845939 0.1014896 0.26395944

By looking above plots and data, it is clearly telling 49th and 50th observations are influencing the P-value. So develop the model by deleting the both observations.

Build the Model:

model2.strtup <- lm(Profit~Spend+Market, data = Strtup[-c(50,49),])

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.090e+04 2.937e+03 17.333 <2e-16 \*\*\*

Spend 7.692e-01 3.485e-02 22.072 <2e-16 \*\*\*

Market 2.514e-02 1.318e-02 1.908 0.0631 .

Residual standard error: 7585 on 43 degrees of freedom

Multiple R-squared: 0.9611, Adjusted R-squared: 0.9575

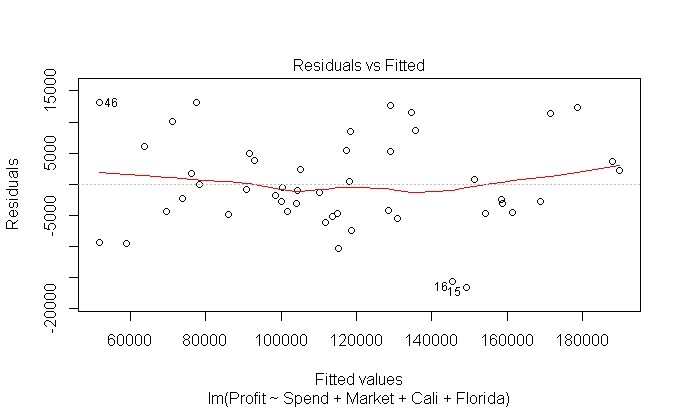
F-statistic: 265.9 on 4 and 43 DF, p-value: < 2.2e-16

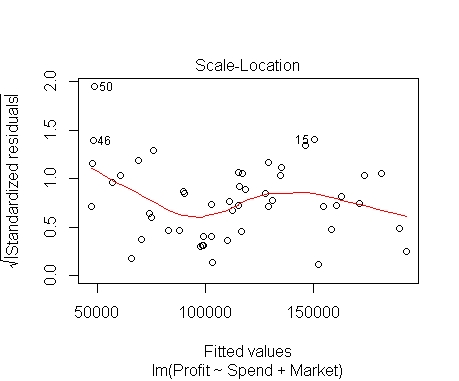
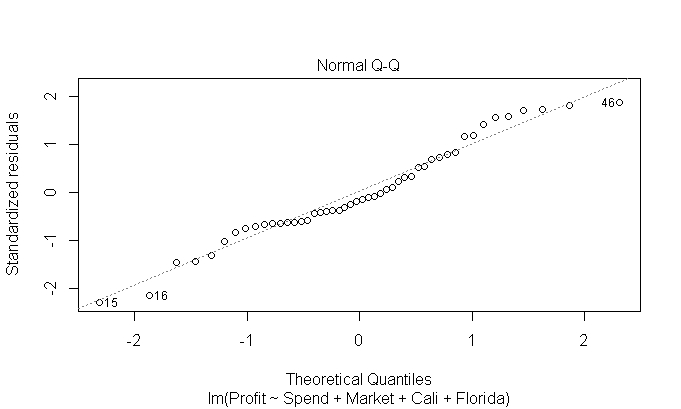
Finally we have good R-squared and Adjusted R2 values are very less difference and have accepted value and Probability of accepted value is lesser than accepted value. So this model is final mode.

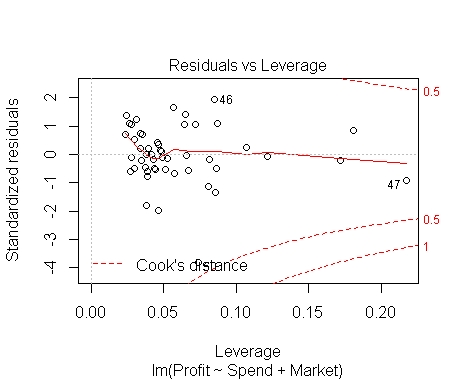
Now, this model has to follow the all LINE assumptions.

**Validating the Line Assumptions:**

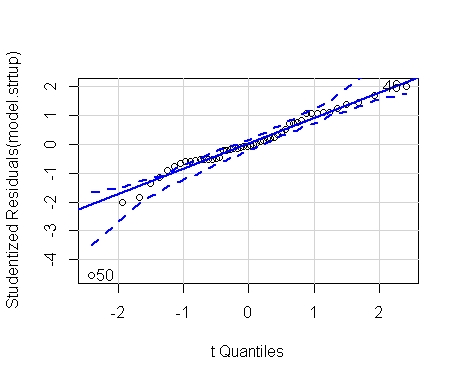
1. Co linearity Checked.
2. Residual Plot : Errors are Identical and has constant variation, so no Homoscedasticity problem.
3. Norm Q-Q Plot : Errors are normally distributed. That mean error has linearity.
4. Standardized Residual :







qqPlot: Difference in the errors.



**2. Problem statement:** Predict sales of the computer

**Data :**

Price is Predicted Value (Y) and it is continuous.

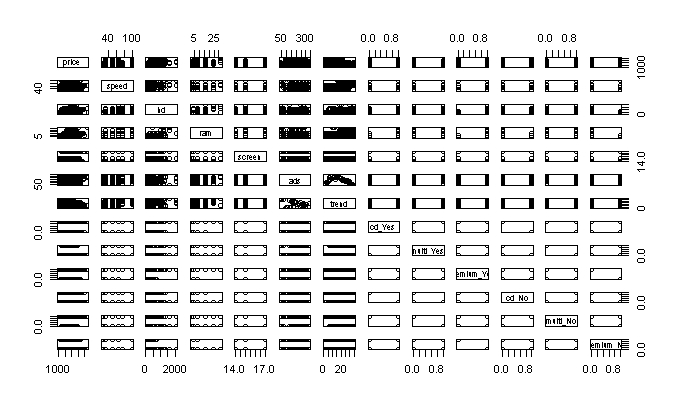
X has number of inputs, in the form of Numeric and Categorical data. Convert Categorical data into Dummy Variables.

Numeric Input data : Speed, hd, ram,ads,trend

Categorical data: cd,multi,premium

**Model Building & Checking the Correlation:**

Plotting the Scatter Diagram:



**Correlation Values:**

Correlation between the **Profit** with other features are values very less except Ram, HD, Speed and Screen.

**Model:**

model.start <- lm(price~speed+hd+ram+screen+cd\_Yes+multi\_Yes+premium\_Yes+ads+trend)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 307.98798 60.35341 5.103 3.44e-07 \*\*\*

speed 9.32028 0.18506 50.364 < 2e-16 \*\*\*

hd 0.78178 0.02761 28.311 < 2e-16 \*\*\*

ram 48.25596 1.06608 45.265 < 2e-16 \*\*\*

screen 123.08904 3.99950 30.776 < 2e-16 \*\*\*

cd\_Yes 60.91671 9.51559 6.402 1.65e-10 \*\*\*

multi\_Yes 104.32382 11.41268 9.141 < 2e-16 \*\*\*

premium\_Yes -509.22473 12.34225 -41.259 < 2e-16 \*\*\*

ads 0.65729 0.05132 12.809 < 2e-16 \*\*\*

trend -51.84958 0.62871 -82.470 < 2e-16 \*\*\*

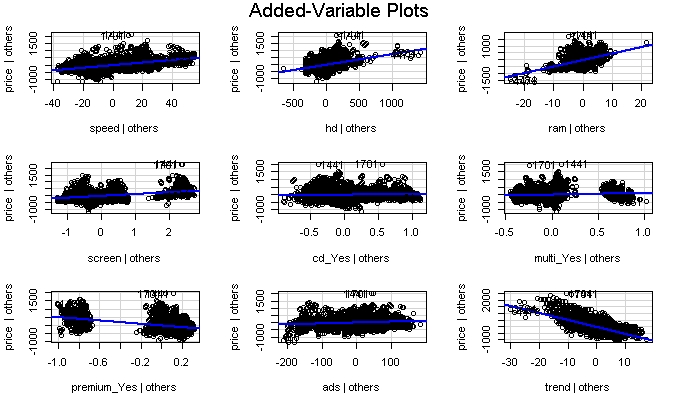
---

Residual standard error: 275.3 on 6249 degrees of freedom

Multiple R-squared: 0.7756, Adjusted R-squared: 0.7752

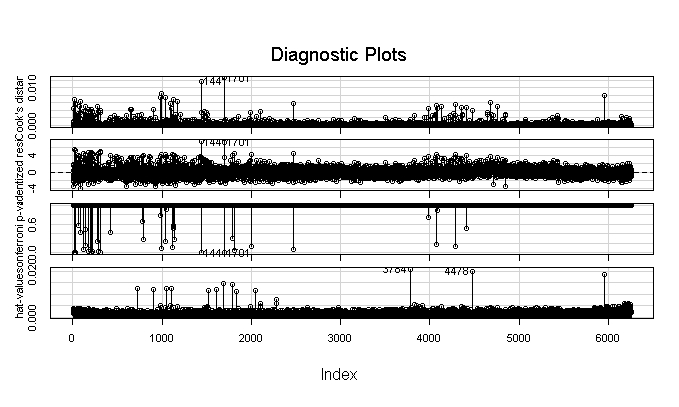
For all the features Probability values are less than accepted value. R-Squared value is not good at the same time not bad. But Probability value is accepted.

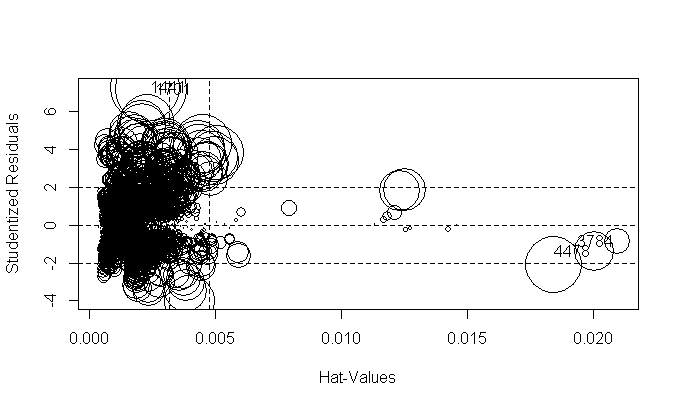
So, there was no co linearity problem. By looking the avPlot, there is two plots are looks some co linearity but their correlation value is very less. So we are neglecting them.



**Checking for influential observation:**

When execute the Influential Index Plot & Influential Plot.

****

****

Look like there are 1441 and 1701 has some influential observations. So we can exclude them and build the model

model1.start <- lm(price~speed+hd+ram+screen+cd\_Yes+multi\_Yes+premium\_Yes+ads+trend, data = Cmpprce[-c(1441,1701),])

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 337.1635 59.9242 5.627 1.92e-08 \*\*\*

speed 9.2992 0.1835 50.664 < 2e-16 \*\*\*

hd 0.7749 0.0274 28.286 < 2e-16 \*\*\*

ram 48.5222 1.0576 45.878 < 2e-16 \*\*\*

screen 121.0926 3.9714 30.492 < 2e-16 \*\*\*

cd\_Yes 60.4964 9.4400 6.409 1.58e-10 \*\*\*

multi\_Yes 104.7703 11.3195 9.256 < 2e-16 \*\*\*

premium\_Yes -509.8352 12.2409 -41.650 < 2e-16 \*\*\*

ads 0.6510 0.0509 12.791 < 2e-16 \*\*\*

trend -51.6496 0.6238 -82.793 < 2e-16 \*\*\*

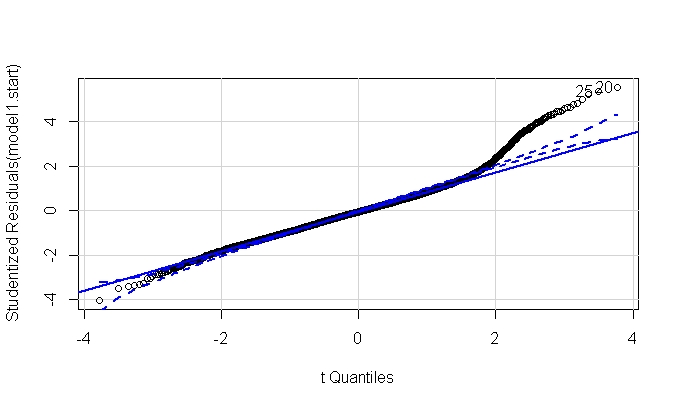
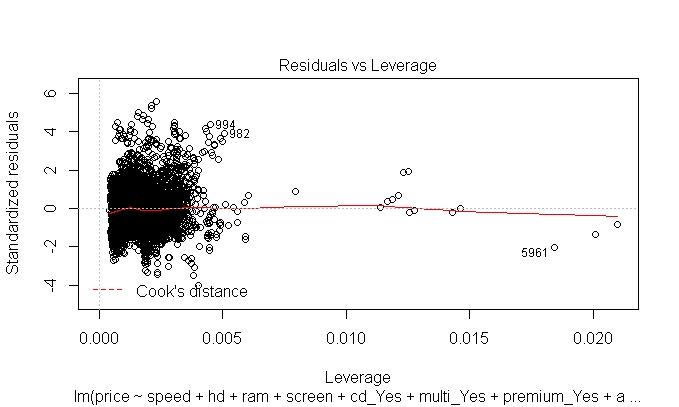
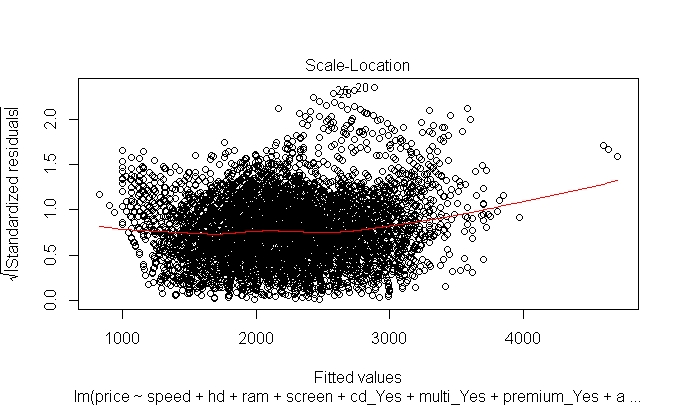
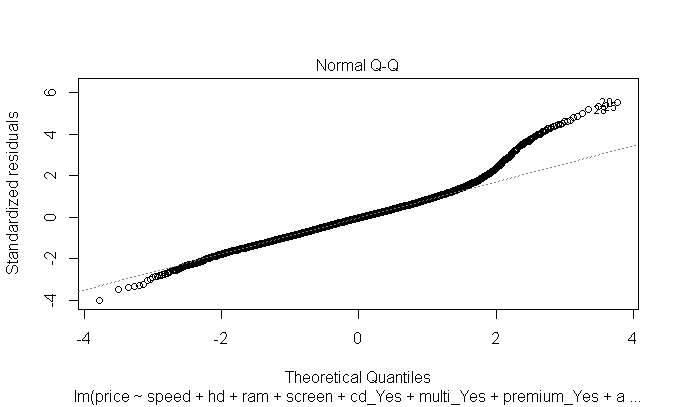
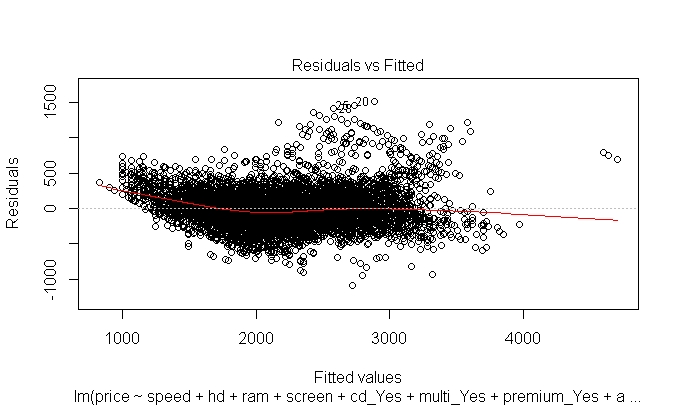
Multiple R-squared: 0.7777, Adjusted R-squared: 0.7774

Finally we have good R-squared and Adjusted R2 values are very less difference and have accepted value and Probability of accepted value is lesser than accepted value. So this model is final mode.

Now, this model has to follow the all LINE assumptions.

**Validating the Line Assumptions:**

1. Co linearity Checked.
2. Residual Plot : Errors are Identical and has constant variation, so no Homoscedasticity problem.
3. Norm Q-Q Plot : Errors are normally distributed. That mean error has linearity.
4. Standardized Residual :



**3. Business Problem:**

Consider only the below columns and prepare a prediction model for predicting Price.

Data :

As per the information consider only following features.

"Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight"

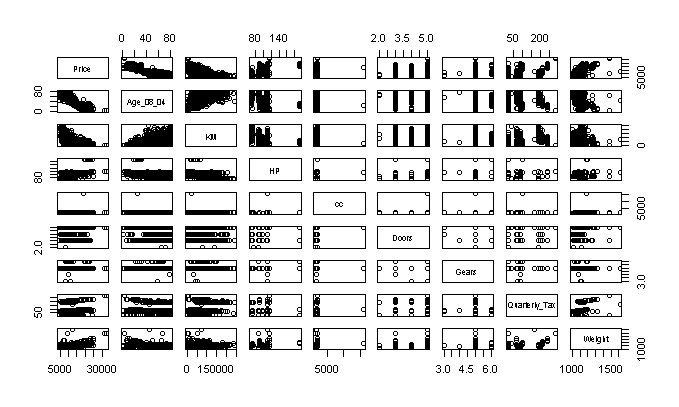
Price will be predicted value and it is continuous data.

Input data are the combination of numeric and categorical data.

So, I’m going with multiple linear regressions.

**Model Building & Checking the Correlation:**

Plotting the Scatter Diagram:



**Correlation Values :**

When we compared the correlation value Age, Km has the negative correlation.HP,Weight and Quarterly tax has the positive correlation.

**Model:**

model.q3 <- lm(Price ~ Age\_08\_04+KM+HP+cc+Doors+Gears+Quarterly\_Tax+Weight,data = Tayot)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.573e+03 1.411e+03 -3.949 8.24e-05 \*\*\*

Age\_08\_04 -1.217e+02 2.616e+00 -46.512 < 2e-16 \*\*\*

KM -2.082e-02 1.252e-03 -16.622 < 2e-16 \*\*\*

HP 3.168e+01 2.818e+00 11.241 < 2e-16 \*\*\*

cc -1.211e-01 9.009e-02 -1.344 0.17909

Doors -1.617e+00 4.001e+01 -0.040 0.96777

Gears 5.943e+02 1.971e+02 3.016 0.00261 \*\*

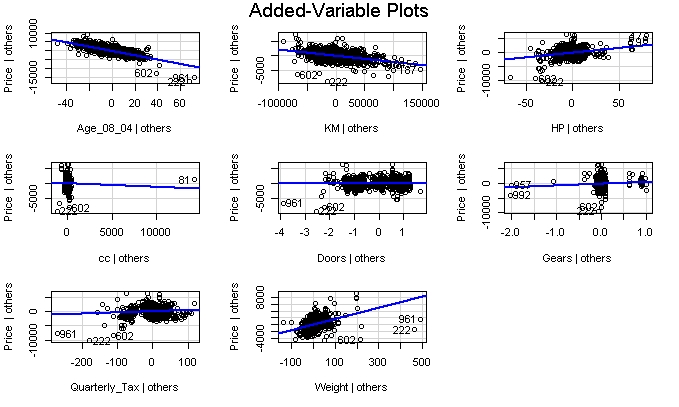
Quarterly\_Tax 3.949e+00 1.310e+00 3.015 0.00262 \*\*

Weight 1.696e+01 1.068e+00 15.880 < 2e-16 \*\*\*

Residual standard error: 1342 on 1427 degrees of freedom

Multiple R-squared: 0.8638, Adjusted R-squared: 0.863

There is good R-Squared value but there will not be a accepted probability for CC and Doors. So we will check for the co linearity by using Variance inflation factor and avPlot.



By looking the avPlot there is co linearity for Doors. So we are removing from model and building a new model.

**Model:**

model2.q3 <- lm(Price ~ Age\_08\_04+KM+HP+cc+Gears+Quarterly\_Tax+Weight,data = Tayot)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.575e+03 1.410e+03 -3.954 8.06e-05 \*\*\*

Age\_08\_04 -1.217e+02 2.615e+00 -46.528 < 2e-16 \*\*\*

KM -2.082e-02 1.251e-03 -16.636 < 2e-16 \*\*\*

HP 3.167e+01 2.810e+00 11.270 < 2e-16 \*\*\*

cc -1.210e-01 9.005e-02 -1.344 0.17909

Gears 5.958e+02 1.934e+02 3.081 0.00210 \*\*

Quarterly\_Tax 3.953e+00 1.306e+00 3.027 0.00251 \*\*

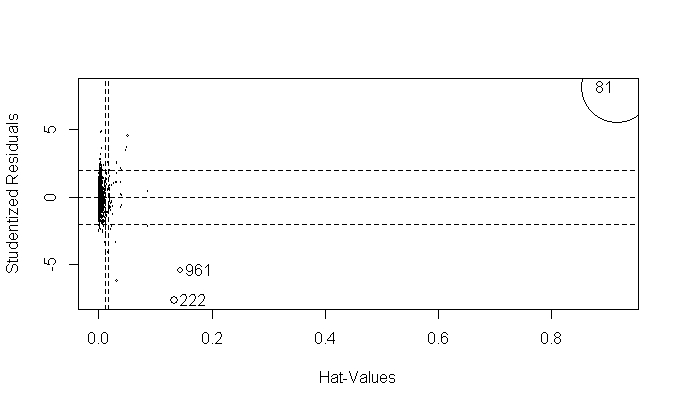
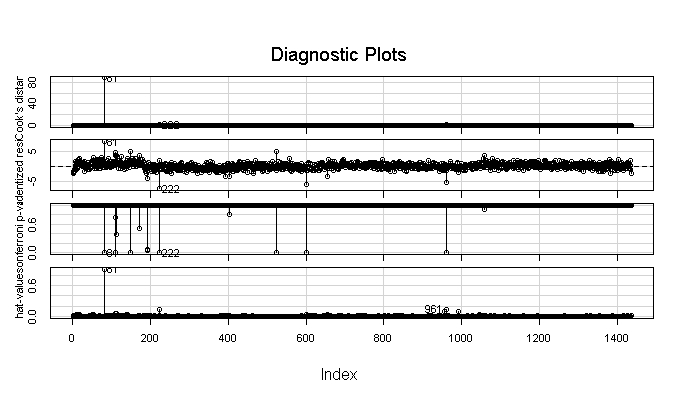
Weight 1.695e+01 1.033e+00 16.401 < 2e-16 \*\*\*

Residual standard error: 1342 on 1428 degrees of freedom

Multiple R-squared: 0.8638, Adjusted R-squared: 0.8631

R-Squared and Adjusted R-squared values are good but still probability value is lesser than accepted value. So we will any influential observations.

Influential Plot:



Clearly the diagram is showing 81 is the influential observation. So eliminating from the data and building a model.

model3.q3 <- lm(Price ~ Age\_08\_04+KM+HP+cc+Gears+Quarterly\_Tax+Weight,data = Tayot[-c(81),])

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.314e+03 1.382e+03 -4.569 5.32e-06 \*\*\*

Age\_08\_04 -1.205e+02 2.561e+00 -47.031 < 2e-16 \*\*\*

KM -1.789e-02 1.275e-03 -14.029 < 2e-16 \*\*\*

HP 3.916e+01 2.898e+00 13.512 < 2e-16 \*\*\*

cc -2.507e+00 3.062e-01 -8.188 5.83e-16 \*\*\*

Gears 5.497e+02 1.892e+02 2.905 0.00373 \*\*

Quarterly\_Tax 9.076e+00 1.424e+00 6.374 2.47e-10 \*\*\*

Weight 1.996e+01 1.076e+00 18.547 < 2e-16 \*\*\*

Residual standard error: 1312 on 1427 degrees of freedom

Multiple R-squared: 0.8693, Adjusted R-squared: 0.8687

Finally we have good R-squared and Adjusted R2 values are very less difference and have accepted value and Probability of accepted value is lesser than accepted value. So this model is final mode.

Now, this model has to follow the all LINE assumptions.

**Validating the Line Assumptions:**

1. Co linearity Checked.
2. Residual Plot : Errors are Identical and has constant variation, so no Homoscedasticity problem.
3. Norm Q-Q Plot : Errors are normally distributed. That mean error has linearity.
4. Standardized Residual :

